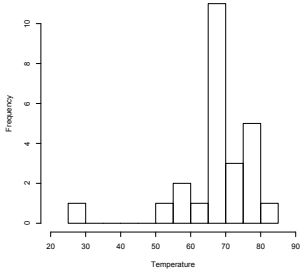
Why are descriptive statistics important?

In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29° F at launch.

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F):
66,70,69,80,68,67,72,70,70,57,63,70,78,67,53,67,75,70,81,76,79,75,76,58,29



The two most important functions of descriptive statistics are:

- Communicate information
- Support reasoning about data

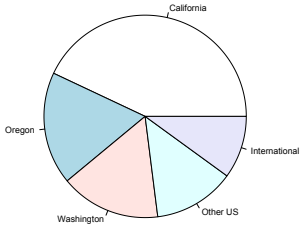
When exploring data of large size, it becomes essential to use summaries.

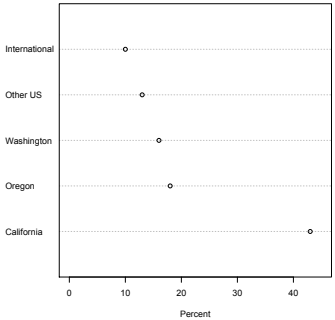
Graphical summaries of data

It is best to use a graphical summary to communicate information, because people prefer to look at pictures rather than at numbers.

There are many ways to visualize data. The nature of the data and the goal of the visualization determine which method to choose.

Pie chart and dot plot



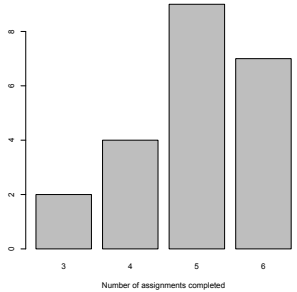


The dot plot makes it easier to compare frequencies of various categories, while the pie chart allows more easily to eyeball what fraction of the total a category corresponds to.

Bar graph

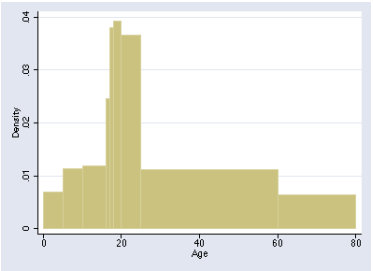
When the data are *quantitative* (i.e. numbers), then they should be put on a number line. This is because the ordering and the distance between the numbers convey important information.

The **bar graph** is essentially a dot plot put on its side.



The histogram

The **histogram** allows to use blocks with different widths.

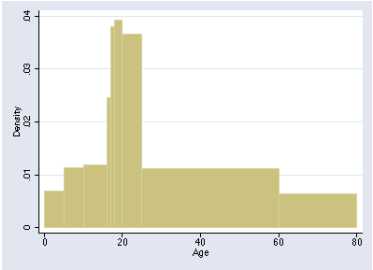


Key point: The areas of the blocks are proportional to frequency.

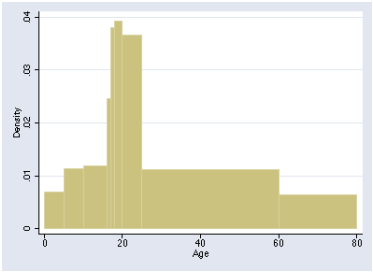
So the percentage falling into a block can be figured without a vertical scale since the total area equals 100%.

But it's helpful to have a vertical scale (*density scale*). Its unit is '% per unit', so in the above example the vertical unit is '% per year'.

The histogram gives two kinds of information about the data:



1. **Density** (crowding): The height of the bar tells how many subjects there are for one unit on the horizontal scale. For example, the highest density is around age 19 as $.04 = 4\%$ of all subjects are age 19. In contrast, only about 0.7% of subjects fall into each one year range for ages 60–80.



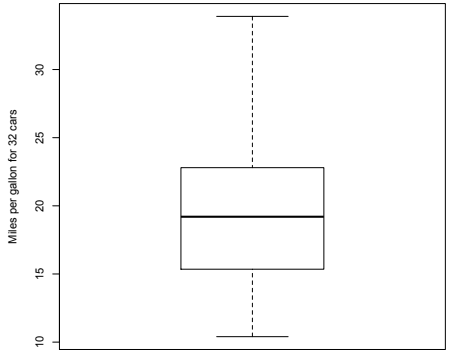
2. **Percentages** (relative frequences): Those are given by

$$\text{area} = \text{height} \times \text{width}.$$

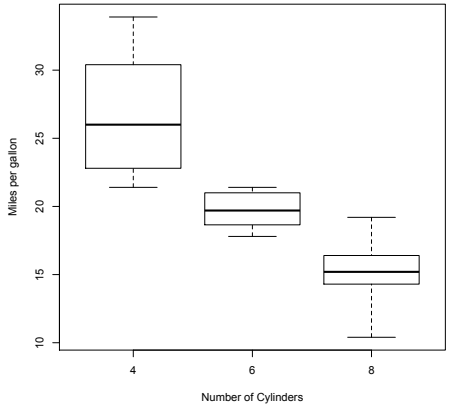
For example, about 14% of all subjects fall into the age range 60–80, because the corresponding area is (20 years) x (0.7 % per year)=14 %. Alternatively, you can find this answer by eyeballing that this area makes up roughly 1/7 of the total area of the histogram, so roughly 1/7=14% of all subjects fall in that range.

The boxplot (box-and-whisker plot)

The **boxplot** depicts five key numbers of the data:

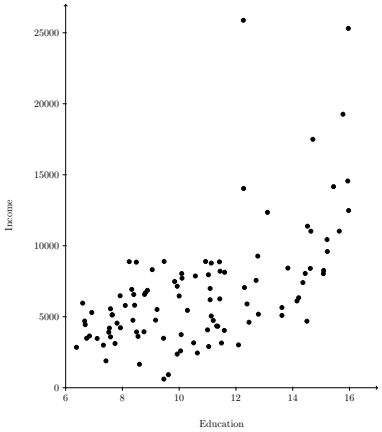


The **boxplot** conveys less information than a histogram, but it takes up less space and so is well suited to compare several datasets:



The scatterplot

The **scatterplot** is used to depict data that come as *pairs*.



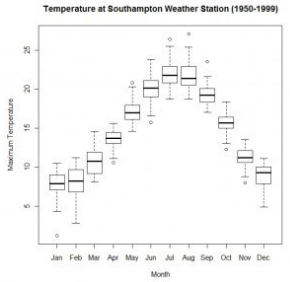
The scatterplot visualizes the relationship between the two variables.

Providing context is important

Statistical analyses typically compare the observed data to a reference. Therefore context is essential for graphical integrity.

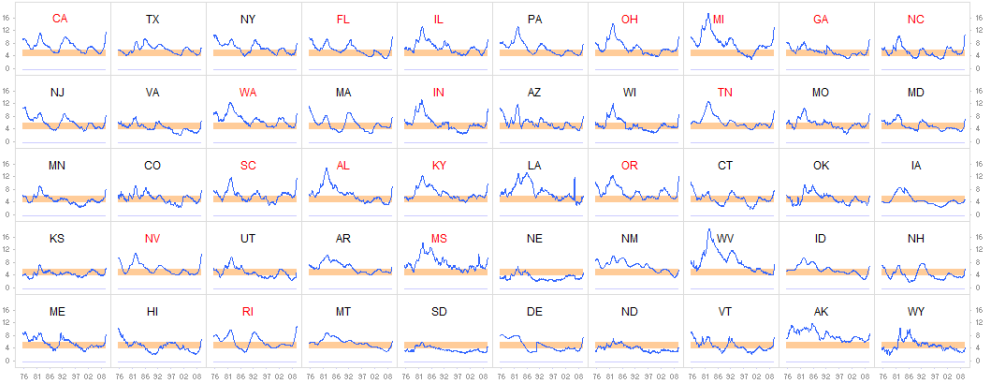
- 'The Visual Display of Quantitative Information' by Edward Tufte (p.74)

One way to provide context is by using *small multiples*. The compact design of the boxplot makes it well suited for this task:



Providing context with small multiples

Monthly Unemployment Rates by State, Jan 1976 - Apr 2009

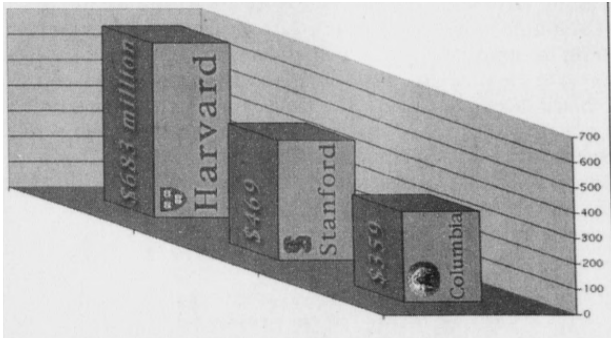

CA TX NY FL IL PA OH MI GA NC
NJ VA WA MA IN AZ WI TN MO MD
MN CO SC AL KY LA OR CT OK IA
KS NV UT AR MS NE NM WV ID NH
ME HI RI MT SD DE ND VT AK WY
16 12 8 4 0
76 81 86 92 97 02 08

Source: Bureau of Labor Statistics

Notes: The orange band denotes a "normal" unemployment rate (4%-6%);
State code in red: unemployment rate in April 2009 is higher than the US average

Pitfalls when visualizing data

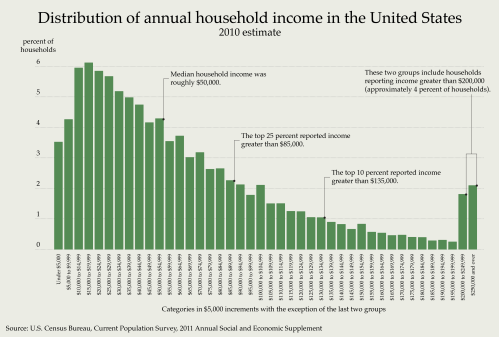
Sophisticated software makes it tempting to produce showy but poor visualizations:



- The 'Ghettysburg Powerpoint Presentation' by Peter Norvig

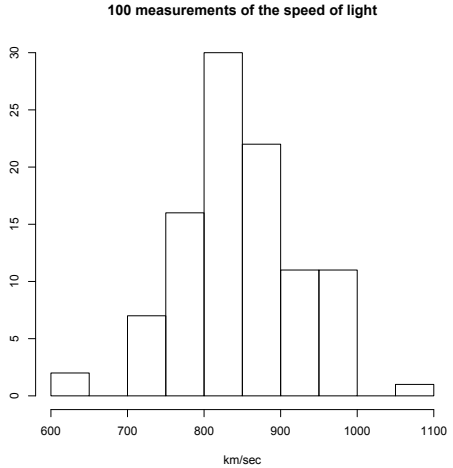
Numerical summary measures

For summerizing data with one number, use the **mean** (=average) or the **median**. The median is the number that is larger than half the data and smaller than the other half.


Distribution of annual household income in the United States
2010 estimate
percent of households
Median household income was roughly $50,000.
The top 25 percent reported income greater than $85,000.
The top 10 percent reported income greater than $135,000.
These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).
Under $5,000
$5,000 to $9,999
$10,000 to $14,999
$15,000 to $19,999
$20,000 to $24,999
$25,000 to $29,999
$30,000 to $34,999
$35,000 to $39,999
$40,000 to $44,999
$45,000 to $49,999
$50,000 to $54,999
$55,000 to $59,999
$60,000 to $64,999
$65,000 to $69,999
$70,000 to $74,999
$75,000 to $79,999
$80,000 to $84,999
$85,000 to $89,999
$90,000 to $94,999
$95,000 to $99,999
$100,000 to $109,999
$110,000 to $119,999
$120,000 to $129,999
$130,000 to $139,999
$140,000 to $149,999
$150,000 to $159,999
$160,000 to $174,999
$175,000 to $199,999
$200,000 to $249,999
$250,000 and over
Categories in $5,000 increments with the exception of the last two groups
Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

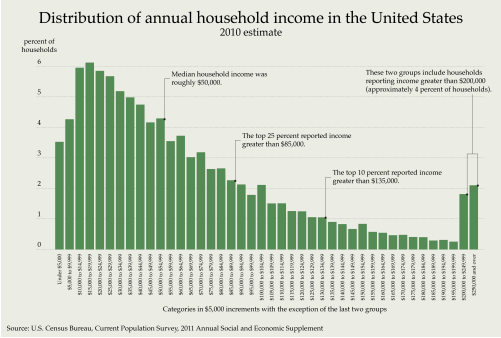
Mean vs. median

Mean and median are the same when the histogram is symmetric.



Mean vs. median

When the histogram is *skewed to the right*, then the mean can be much larger than the median.


Distribution of annual household income in the United States
2010 estimate
percent of households
Median household income was roughly $50,000.
The top 25 percent reported income greater than $85,000.
The top 10 percent reported income greater than $135,000.
These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).
Categories in $5,000 increments with the exception of the last two groups
Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

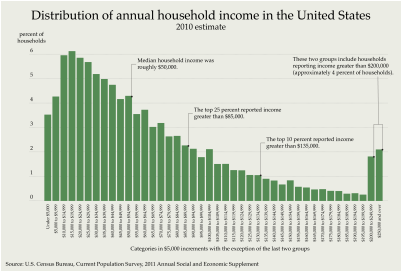
So if the histogram is very skewed, then use the median.

Mean vs. median

If the median sales price of 10 homes is $ 1 million, then we know that 5 homes sold for $ 1 million or more.

If we are told that the average sale price is $ 1 million, then we can't draw such a conclusion:

Percentiles


Distribution of annual household income in the United States
2010 estimate
percent of households
Median household income was roughly $50,000.
These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).
The top 25 percent reported income greater than $85,000.
The top 10 percent reported income greater than $135,000.
Categories in $5,000 increments with the exception of the last two groups
Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

The 90th percentile of incomes is $ 135,000: 90% of households report an income of $ 135,000 or less, 10% report more.

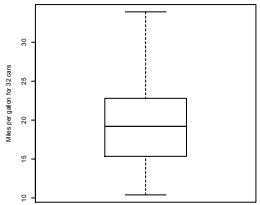
The 75th percentile is called **3rd quartile**: $ 85,000

The 50th percentile is the **median**: $ 50,000

The 25th percentile is called **1st quartile**.

Five-number summary

Recall that the boxplot gives a **five-number summary** of the data:
the smallest number, 1st quartile, median, 3rd quartile, largest number.



The **interquartile range** = 3rd quartile − 1st quartile.
It measures how spread out the data are.

The standard deviation

A more commonly used measure of spread is the **standard deviation**.

$\bar{x}$ stands for the average of the numbers $x_1, \ldots, x_n$.

The standard deviation of these numbers is

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{or} \quad \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The two numbers $\bar{x}$ and s are often used to summarize data. Both are sensitive to a few large or small data.

If that is a concern, use the median and the interquartile range.

What is statistical inference?

What percentage of voters approve of the way the U.S. President is handling his job?

This is difficult to determine exactly as there are more than 250 million people of voting age in the U.S.

But it's not difficult to *estimate* this percentage quite well:

Sample 1,000 (say) voters at random. Then use the approval percentage among those voters as an estimate for the approval percentage of all voters.

What is statistical inference?

Population: the entire group of subjects about which we want information
all U.S. voters

Parameter: the quantity about the population we are interested in
approval percentage among all U.S. voters

Sample: the part of the population from which we collect information
the 1,000 voters selected at random

Statistic (estimate): the quantity we are interested in as measured in the sample
approval percentage among the sampled voters

Key point: even a relatively small sample (100 or 1,000) will produce an estimate that is close to the parameter of a very large population of 250 million subjects. This is the reason why statistics is so powerful.

Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

selection bias: a sample of convenience makes it more likely to sample certain subjects than others

non-response bias: parents are less likely to answer a survey request at 6 pm because they are busy with children and dinner

voluntary response bias: websites that post reviews of businesses are more likely to get responses from customers who had very bad or very good experiences

Sampling designs

The best methods for sampling use chance in a planned way:

a **simple random sample** selects subjects at random without replacement

a **stratified random sample** divides the population into groups of similar subjects called *strata* (e.g. urban, suburban, and rural voters). Then one chooses a simple random sample in each stratum and combines these.

Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} \;=\; \text{parameter} + \text{bias} + \text{chance error}$$

The chance error (sampling error) will get smaller as the sample size gets bigger. Moreover, we can compute how large the chance error will be.

This is not the case for the bias (systematic error):
Increasing the sample size just repeats the error on a larger scale, and typically we don't know how large the bias is.

Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- This means that there is an **association** between red meat consumption and cancer: there is a link between these two.
- But this does **not** mean that eating red meat *causes* cancer: people who don't eat red meat are known to exercise more and drink less alcohol, and it could be the latter two issues that cause the difference in cancer rates.

This is an **observational study**: It measures outcomes of interest and this can be used to establish association.

But **association is not causation**, because there may be **confounding factors** such as exercise that are associated both with red meat consumption and cancer.

Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- The **subjects** are assigned into treatment and control groups at **random**.
- When possible, subjects in the control group get a **placebo**: it resembles the treatment but is neutral. Assigning a placebo makes sure that both groups are equally affected by the **placebo effect**: the idea of being treated may have an effect by itself.
- The experiment is **double-blind**: neither the subjects nor the evaluators know the assignments to treatment and control.

The placebo effect

The placebo effect is still not fully understood and is one of the most interesting phenomena in science.

'The weird power of the placebo effect, explained' by Brian Resnick (7/7/2017)

The logic of randomized controlled experiments

Randomization serves two purposes:

- It makes the treatment group similar to the control group. Therefore influences other than the treatment operate equally on both groups, apart from differences due to chance.
- It allows to assess how relevant the treatment effect is, by calculating the size of chance effects when comparing the outcomes in the two groups (see later).

What is probability?

In 2015 there were about 4 million babies born in the US, and 48.8% of the newborns were girls.

We write: P(newborn is a girl)$= 48.8\%$.

The probability of an event is defined as the proportion of times this event occurs in many repetitions.

This is the standard definition of probility. It requires that it is possible to repeat this chance experiment many times.

While interned during the Second World War, John Kerrich tossed a coin 10,000 times and observed 5,067 tosses resulting in heads.

What is probability?

The long-run interpretation of probability can make it difficult to interpret it for single event:
'What I was wrong about this year' by David Leonhardt (12/24/2017)

Sometimes people use a different interpretation:
'The probability that my best friend calls today is 30%'.

Such a 'subjective probability' is not based on experiments, and different people may assign different subjective probabilities to the same event.

Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that P(newborn is a girl)$= 48.8\%$.
Therefore P(newborn is a boy)$= 51.2\%$.

More formally, write A for an event, such as A='newborn is a girl'.

Complement rule: P(A does not occur) = 1−P(A)

Now let's look at a different chance experiment: Rolling a die.

Since each of its six faces is equally likely to come up, each has probability 1/6.

Rule for equally likely outcomes: If there are n possible outcomes and they are equally likely, then $\mathrm{P(A)} = \frac{\text{number of outcomes in A}}{n}$

Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

As an example, we roll a die twice.
A= ⚀ on first roll, B= ⚅ on first roll, C= ⚀ on second roll

A and B are mutually exclusive, but A and C are not.

Addition rule: If A and B are mutually exclusive, then

$$P(\text{A or B}) = P(\text{A}) + P(\text{B})$$

Two events are *independent* if knowing that one occurs does not change the probability that the other occurs.
B and C are independent, but A and B are not.

Multiplication rule: If A and B are independent, then

$$P(\text{A and B}) = P(\text{A})\,P(\text{B})$$

Four basic rules

Roll a die three times. What is P(at least one ⚅) ?

We could write 'at least one ⚅' as follows:

⚅ on the first roll or ⚅ on the second roll or ⚅ on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

P(at least one ⚅ in three rolls)

$= 1 - $P(no ⚅ in three rolls)

$= 1 - $P((no ⚅ in first roll) and (no ⚅ in second roll) and (no ⚅ in third roll))

$= 1 - $P(no ⚅ in first roll)$\times$ P(no ⚅ in second roll)$\times$P(no ⚅ in third roll)

$= 1 - \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$

$= 41.1\%$

Conditional probability

Spam e-mail has a higher chance to contain the word 'money' than ham e-mail:

P('money' in e-mail | spam)$= 8\%$, P('money' in e-mail | ham)$= 1\%$.

The *conditional probability of B given A* is

$$\mathrm{P(B|A)} \;=\; \frac{\mathrm{P(A \text{ and } B)}}{\mathrm{P(A)}}$$

General multiplication rule: $\mathrm{P(A \text{ and } B) = P(A)\,P(B|A)}$

In the special case where A and B are independent: $\mathrm{P(A \text{ and } B) = P(A)\,P(B)}$

Conditional probability

Computing probabilities by total enumeration:
P(spam) = 20%. What is the probability that 'money' appears in an e-mail?

From data we know P(money | spam)$= 8\%$, P(money | ham)$= 1\%$.

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

money appears and e-mail is spam or money appears and e-mail is ham

P(money appears)

=P(money and spam) + P(money and ham)

=P(money | spam) P(spam) + P(money | ham) P(ham)

$= 0.08 \times 0.2 + 0.01 \times 0.8$

$= 2.4\%$

Bayes' rule

From data we know P(money appears in e-mail | e-mail is spam)$= 8\%$, but what we need to build a spam filter is P(e-mail is spam | money appears in e-mail).

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(B \text{ and } A)}{P(A)} = \frac{P(A|B)\ P(B)}{P(A)}$$

$$P(\text{spam} \mid \text{money}) = \frac{P(\text{money} \mid \text{spam})\ P(\text{spam})}{P(\text{money})} = \frac{0.08 \times 0.2}{0.024} = 67\%$$

Bayes' rule:

$$P(B|A) = \frac{P(A|B)\ P(B)}{P(A)}$$

$$= \frac{P(A|B)\ P(B)}{P(A|B)P(B) + P(A|\text{not } B)\ P(\text{not } B)}$$

Bayesian analysis

The spam filter classifies e-mail as spam via a *Bayesian analysis*:

- Before examining the e-mail, there is a *prior probability* of 20% that it is spam.
- After examining the e-mail for certain keywords such as 'money', the filter updates this prior probability using Bayes' rule to arrive at the *posterior probability* that the e-mail is spam.

Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know P(D)$= 1\%$, P(+|D)$= 95\%$, P(+|no D)$= 2\%$.

$$\text{Want P(D}|+) = \frac{\text{P}(+|\text{D})\ \text{P(D)}}{\text{P}(+)}$$

$$= \frac{\text{P}(+|\text{D})\ \text{P(D)}}{\text{P}(+|\text{D})\ \text{P(D)} + \text{P}(+|\text{no D})\ \text{P(no D)}}$$

$$= \frac{0.95\times 0.01}{0.95\times 0.01 + 0.02\times 0.99} = 32.4\%$$

Case study: Warner's randomized response model

What percentage of students have cheated during an exam in college?

Problem: Students may be too embarrassed to answer truthfully.

Randomization comes to the rescue:

We do a survey that first instructs students to toss a coin twice. If the student gets 'tails' on the first toss, then the student has to answer question 1, otherwise the student answers question 2.

Q1: Have you ever cheated on an exam in college?

Q2: Did you get 'tails' on the second toss?

So the answer will be partly random: We don't know whether a 'yes' answer is due to the student cheating or to getting tails on the second toss. This should put the student at ease to answer truthfully.

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

$$\begin{aligned} \text{P(yes)} &= \text{P(yes and Q1)} + \text{P(yes and Q2)} \\ &= \text{P(yes} \mid \text{Q1)}\ \text{P(Q1)} + \text{P(yes} \mid \text{Q2)}\ \text{P(Q2)} \end{aligned}$$

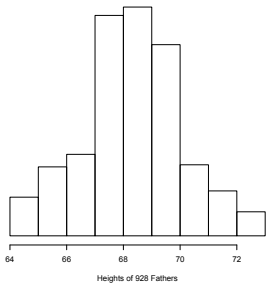
$$\text{Solve for P(yes} \mid \text{Q1)} = \frac{\text{P(yes)} - \text{P(yes} \mid \text{Q2)}\ \text{P(Q2)}}{\text{P(Q1)}}$$

In one survey, 27 students answered 'yes' and 30 answered 'no'.

So we estimate $\text{P(yes)} = \frac{27}{27+30} = 47\%$ and get $\text{P(yes} \mid \text{Q1)} = \frac{0.47 - 0.5 \times 0.5}{0.5} = 44\%$

The normal curve

Many data have histograms that look bell-shaped, e.g. heights, weights, IQ scores:



'The data follow the normal curve.'

But remember that some data have histograms that look quite different, e.g. incomes, house prices.

The empirical rule

If the data follow the normal curve, then

- about 2/3 (68%) of the data fall within one standard deviation of the mean
- about 95% fall within 2 standard deviations of the mean
- about 99.7% fall within 3 standard deviations of the mean

Galton's measurements of heights of fathers have $\bar{x} = 68.3$ in and $s = 1.8$ in. Therefore about 95% of all heights are between 68.3 in -2×1.8 in $= 64.7$ in and 68.3 in $+2 \times 1.8$ in $= 71.9$ in.

The empirical rule

Recall that in a histogram, percentages are given by areas:

Standardizing data

A normal curve is determined by $\bar{x}$ and s: If the data follow the normal curve, then knowing $\bar{x}$ and s means knowing the whole histogram.

To compute areas under the normal curve, we first **standardize** the data by subtracting off $\bar{x}$ and then dividing by s:

$$z = \frac{\text{height} - \bar{x}}{s}$$

z is called the **standardized value** or **z-score**.

z has no unit (height, $\bar{x}$ and s all have the unit 'inches')

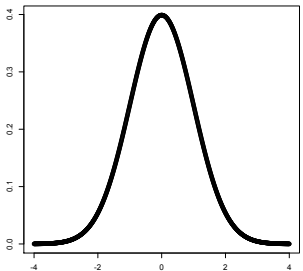
For example, $z = 2$ means the height is 2 standard deviations above average.

$z = -1.5$ means the height is 1.5 standard deviations *below* average.

The standard normal curve

Standardized data have mean 0 and standard deviation equal to 1 – this is the point of standardizing.

Fathers' heights follow the normal curve with $\bar{x} = 68.3$ in and $s = 1.8$ in. Therefore the standardized values follow the **standard normal curve** with mean 0 and standard deviation 1.



This curve is given by the function

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Normal approximation

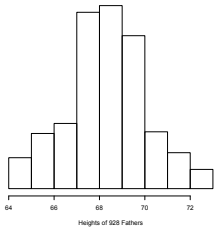
Finding areas under the normal curve is called **normal approximation**.

What percentage of fathers have heights between 67.4 in and 71.9 in?

1. Standardize:

$$\frac{67.4 \text{ in} - 68.3 \text{ in}}{1.8 \text{ in}} = -0.5 \qquad \frac{71.9 \text{ in} - 68.3 \text{ in}}{1.8 \text{ in}} = 2$$

2. Mark the area under the normal curve:



Normal approximation

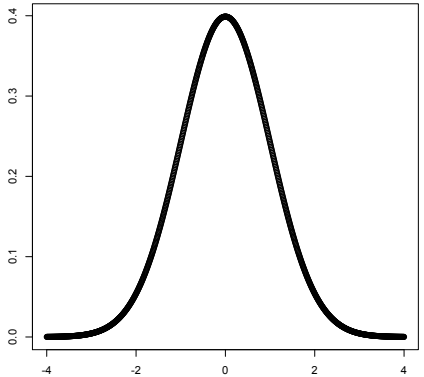
3. Write the desired area in a form that can be computed by software or looked up in a table:

   Typically we can look up the area to the left of a given value.

4. Use software or a table to find these values: $97.7\% - 30.9\% = 66.8\%$

Normal approximation

The empirical rule is a special case of normal approximation:



Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Recall $z = \frac{\text{height} - \bar{x}}{s}$.

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Recall $z = \frac{\text{height} - \bar{x}}{s}$. Solve for height$= \bar{x} + zs$

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Recall $z = \frac{\text{height} - \bar{x}}{s}$. Solve for height$= \bar{x} + zs$

Or: $z = -0.52$ means that the height is 0.52 standard deviations below average.

Computing percentiles for normal data

What is the 30th percentile of the fathers' heights?

From software or from a normal table: $z = -0.52$

Recall $z = \frac{\text{height} - \bar{x}}{s}$. Solve for height$= \bar{x} + zs$

Or: $z = -0.52$ means that the height is 0.52 standard deviations below average. So the height is $\bar{x} - 0.52s = 68.3$ in $-(0.52)(1.8 \text{ in}) = 67.4$ in.

The binomial setting

We saw that for a newborn baby, there is a 49% chance that it is a girl.

What are the chances that 2 out of 3 newborns are girls?

We can compute this by listing all the possibilities (*total enumeration*):

$$\begin{aligned}
\text{P(2 out of 3 are girls)} &= \text{P(GGB or GBG or BGG)} \\
&= \text{P(GGB)} + \text{P(GBG)} + \text{P(BGG)} && \text{addition rule} \\
&= \text{P(G) P(G) P(B)} + \text{P(G) P(B) P(G)} + \ldots && \text{multiplication rule} \\
&= 3 \times (0.49)(0.49)(0.51)
\end{aligned}$$

'3' counts the number of ways one can arrange two G and one B

The binomial setting

This is an example of the **binomial setting**:

- There are $n = 3$ *independent* repetitions of an experiment.
- Each of these experiments has two possible outcomes (which are generically called 'success' and 'failure').
- The probability of success $p = 49\%$ is the same in each experiment.

The binomial coefficient

What about P(2 out of 5 newborns are girls) ?

In principle, we can compute this in the same way. However, there are now 10 possibilities to arrange 2 girls among 5 newborns:

GGBBB, GBGBB, GBBGB,...

The number of possibilities grows very quickly as n gets larger. Fortunately, there is a formula for it:

The **binomial coefficient** counts the number of ways one can arrange k successes in n experiments:

$$\frac{n!}{k!(n-k)!} \qquad \text{where } n! = 1 \times 2 \times 3 \times \ldots \times n$$

$0! = 1$. We had $n = 3$ and $k = 2$, so $\frac{3!}{2!1!} = \frac{1\times2\times3}{1\times2\times1} = 3$.

The binomial formula

Applying this coefficient in the binomial setting gives:

P(k successes in n experiments) $= \frac{n!}{k!(n-k)!}\ p^k(1-p)^{n-k}$

This is the **binomial probability**.

The binomial formula

You play an online game 10 times. Each time there are three possible outcomes: P(win a big prize) $= 10\%$, P(win a small prize) $= 20\%$, P(win nothing) $= 70\%$.

What is P(win two small prizes) ?

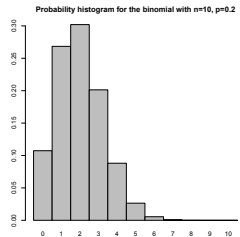
Random variables

The outcomes of the 10 experiments are due to chance, so the number of successes is random: One set of 10 experiments might result in 4 successes, another set might result in 7 successes.

$X =$'number of successes' is called a **random variable**.

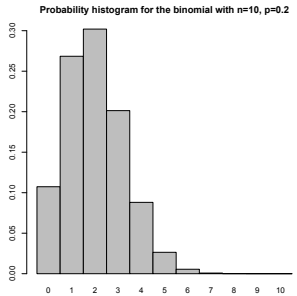
$\mathrm{P}(X = 2) = 30.2\%$. X has the **binomial distribution**.

We can visualize the probabilities of the various outcomes of X with a **probability histogram**:



The probability histogram

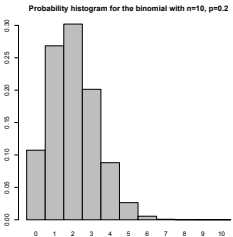
We can visualize the probabilities of the various outcomes of X with a **probability histogram**:

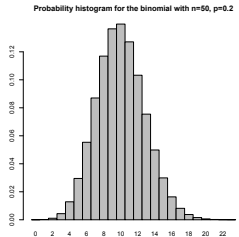


A histogram of data gives percentages for observed data. In contrast, a probability histogram is a theoretical construct: it visualizes probabilities rather than data that have been empirically observed.

Normal approximation to the binomial

As the number of experiments n gets larger, the probability histogram of the binomial distribution looks more and more similar to the normal curve:





In fact, we can approximate binomial probabilities using **normal approximation**: to standardize, subtract off np and then divide by $\sqrt{np(1-p)}$.

Normal approximation to the binomial

We can approximate binomial probabilities using **normal approximation**: to standardize, subtract off np and then divide by $\sqrt{np(1-p)}$.

In the previous example, we had $p =$ P(win a small prize) $= 0.2$.
Play $n = 50$ times. What is P(at most 12 small prizes) ?

Sampling without replacement

A **simple random sample** selects subjects without replacement.

This is not the binomial setting, because p changes after a subject has been removed.

But if the population is much larger than the sample, then sampling with replacement is about the same as sampling without replacement.

Then the number of successes will have approximately the binomial distribution (and so it will approximately follow the normal curve.)

Parameter and statistic

What is the average height of adult men in the US?

This is difficult to evaluate as there are 120 million adult men, but it can be *estimated* quite well with a relatively small sample.

The **population** consists of all adult men in the US (about 120 million).

A **parameter** is a quantity of interest about the population: the population average μ, or the population standard deviation σ.

A **statistic (estimate)** is the quantity of interest as measured in the sample: the sample average $\bar{x}$, or the sample standard deviation s.

The expected value

If we sample an adult male at random, then we expect his height to be around the population average μ, give or take about one standard deviation σ.

The **expected value** of one random draw is the population average μ.

How about $\bar{x}_n$, the average of n draws?

The **expected value of the sample average, E($\bar{x}_n$)**, is the population average μ.

But remember that $\bar{x}_n$ is a random variable because sampling is a random process. So $\bar{x}_n$ won't be exactly equal to $\mu = 69.3$ in: We might get, say, $\bar{x}_n = 70.1$ in. Taking another sample of size n might result in $\bar{x}_n = 69.1$ in.

How far off from μ will $\bar{x}_n$ be?

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

The standard error for the sample average

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

So the SE for a statistic plays the same role that the standard deviation σ plays for one observation drawn at random.

The **square root law** is key for statistical inference:

$$\text{SE}\ (\bar{x}_n)\ =\ \frac{\sigma}{\sqrt{n}}$$

The importance of the square root law is twofold:

- It shows that the SE becomes smaller if we use a larger sample size n. We can use the formula to determine what sample size is required for a desired accuracy.
- The formula for the standard error **does not depend on the size of the population**, only on the size of the sample.

Expected value and standard error for the sum

What if we are interested in the sum of the n draws, S_n, rather than the average $\bar{x}_n$?

The sum and the average are related by $S_n = n\bar{x}_n$.
Both the expected value and the standard error can likewise be obtained by multiplying by n, therefore

$$\mathsf{E}(S_n) = n\mu \qquad \mathsf{SE}(S_n) = \sqrt{n}\sigma$$

So the variability of the sum of n draws *increases* at the rate $\sqrt{n}$.

Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The 'percentage of likely voters' is an average. This becomes clear by using the **framework for counting and classifying**:

- The population consists of all likely voters (about 140 million).
- Each likely voter falls into one of two categories: approve or not approve.
- Put the label '1' on each likely voter who approves, and '0' on each who doesn't.
- Then the number of likely voters who approve equals the sum of all 140 million labels.
-
- The percentage of likely voters who approve is the percentage of 1s among the labels.

Expected value and standard error for percentages

In a sample of n likely voters

- the number of voters in the sample who are approving is the sum S_n of the draws
- the percentage of voters approving is the percentage of 1s, which is $\frac{S_n}{n} \times 100\% = \bar{x}_n \times 100\%$

Therefore

$$\text{E(percentage of 1s)} = \mu \times 100\% \qquad \text{SE(percentage of 1s)} = \frac{\sigma}{\sqrt{n}} \times 100\%$$

where μ is the population average (=proportion of 1s) and σ is the standard deviation of the population of 0s and 1s.

All of the above formulas are for sampling with replacement. They are still approximately true when sampling without replacement if the sample size is much smaller than the size of the population.

Expected value and standard error when simulating

All of the above formulas are also true when the data are **simulated**, i.e. generated according to a probability histogram.
What are μ and σ in that case?

If the random variable X that is simulated has K possible outcomes $x_1, \ldots, x_K$, then

$$\mu = \sum_{i=1}^{K} x_i \mathsf{P}(X = x_i) \qquad \sigma^2 = \sum_{i=1}^{K} (x_i - \mu)^2 \mathsf{P}(X = x_i)$$

If the random variable X has a density f, such as when X follows the normal curve, then

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \qquad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Applying the square root law

Toss a coin 100 times. How many ‘tails’ do you expect to see? Give or take how many?

The sampling distribution

Toss a coin 100 times. The number of tails has the following possible outcomes: $0, 1, 2, \ldots, 100$.

How likely is each outcome?

The number of tails has the binomial distribution with $n = 100$ and $p = 0.5$. ('success' = coin lands tails)

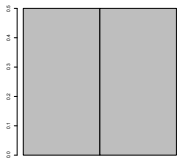
So if the statistic of interest is S_n='number of tails', then S_n is a random variable whose probability histogram is given by the binomial distribution. This is called the **sampling distribution** of the statistic S_n.

The sampling distribution of S_n provides more detailed information about the chance properties of S_n than the summary numbers given by the expected value and the standard error.

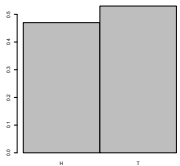
There are three histograms

The chance process of tossing a coin 100 times comes with three different histograms:

1. The probability histogram for producing the data:

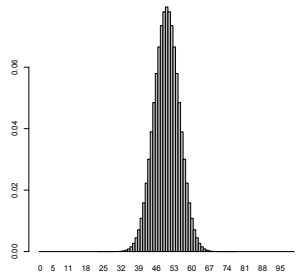


2. The histogram of the 100 observed tosses. This is an empirical histogram of real data:



There are three histograms

3. The probability histogram of the statistic S_{100} = 'number of tails', which shows the sampling distribution of S_{100}:



When doing statistical inference it is important to carefully distinguish these three histograms.

The law of large numbers

The square root law says that SE($\bar{x}_n$), the standard error of the sample mean, goes to zero as the sample size increases.

Therefore the sample mean $\bar{x}_n$ will likely be close to its expected value μ if the sample size is large. This is the **law of large numbers**.

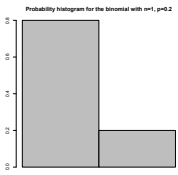
Keep in mind that the law of large numbers applies

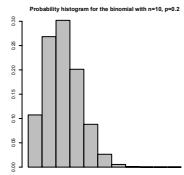
- for averages and therefore also for percentages, but not for sums as their SE *increases*
- for sampling with replacement from a population, or for simulating data from a probability histogram

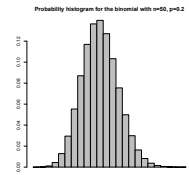
More advanced versions of the law of large numbers state that the empirical histogram of the data (the histogram in 2. in the previous section) will be close to the probability histogram in 1. if the sample size is large.

The central limit theorem

Recall the online game where you win a small prize with probability 0.2. We looked at the random variable $X =$'number of small prizes' in n gambles and found that X has the binomial distribution with that n and $p = 0.2$.





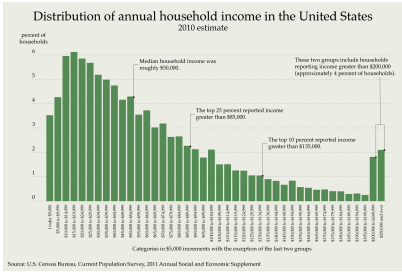


As n gets large, the probability histogram looks more and more similar to the normal curve. This is an example of the **central limit theorem**:

When sampling with replacement and n is large, then the sampling distribution of the sample average (or sum or percentage) approximately follows the normal curve. To standardize, subtract off the expected value of the statistic, then divide by its SE.

The central limit theorem

The key point of the theorem is that we know that the sampling distribution of the statistic is normal *no matter what the population histogram is*:


Distribution of annual household income in the United States
2010 estimate
percent of households
Median household income was roughly $50,000.
The top 25 percent reported income greater than $85,000.
The top 10 percent reported income greater than $135,000.
Those two groups include households reporting income greater than $200,000 (approximately 4 percent of households).
Categories in $5,000 increments with the exception of the last two groups
Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

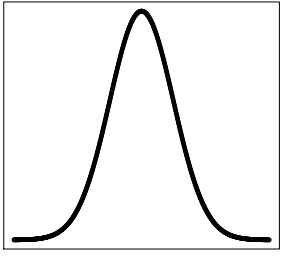
$$\mu = \$\,67,000$$
$$\sigma = \$\,38,000$$

If we sample n incomes at random, then the sample average $\bar{x}_n$ follows the normal curve centered at $\mathsf{E}(\bar{x}_n) = \mu = \$\,67,000$ and with its spread given by $\mathsf{SE}(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$\,38{,}000}{\sqrt{n}}$.

The central limit theorem

If we sample n incomes at random, then the sample average $\bar{x}_n$ follows the normal curve centered at $\mathsf{E}(\bar{x}_n) = \mu = \$\,67,000$ and with its spread given by $\mathsf{SE}(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$\,38{,}000}{\sqrt{n}}$.

For example, if we sample 100 incomes, then by the empirical rule there is about a 16% chance that $\bar{x}_n$ is larger than \$ 70,800:



The central limit theorem

In the example about online gambling, we had $X =$'number of small prizes' in n gambles.

Since we are counting the number of small prizes, we use a label for each gamble which shows '1' if a small prize is won and '0' otherwise.

Then X equals the sum of these labels and so the central limit theorem applies.

Using the formulas for μ and σ in the case of simulations we find $\mu = p$ and $\sigma = \sqrt{p(1-p)}$.

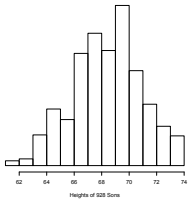
Therefore we standardize the sum X with the expected value np and $\mathsf{SE}(X) = \sqrt{np(1-p)}$.

When does the central limit theorem apply?

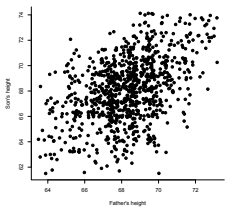
For the normal approximation to work, the key requirements are:

- We sample with replacement, or we simulate independent random variables from the same distribution.
- The statistic of interest is a sum (averages and percentages are sums in disguise).
- The sample size is large enough: the more skewed the population histogram is, the larger the required sample size n.
  (if there is no strong skewness then $n \geq 15$ is sufficient)

Prediction is a key task of statistics

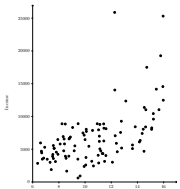


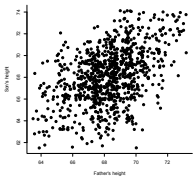
Predict the height of a son who is chosen at random from 928 sons. The average height of sons, 68.1 in, is the 'best' predictor.



Predict the height of a son whose father is 72 in tall. This additional information about the father should allow us to make a better prediction. Regression does just that.

The correlation coefficient





The scatterplot visualizes the relationship between two quantitative variables. It may have a direction (sloping up or down), form (a scatter that clusters around a line is called *linear*) and strength (how closely do the points follow the form?).

If the form is linear, then a good measure of strength is the **correlation coefficient r**: Our data are $(x_i, y_i),\ i = 1, \ldots, n.$

$$r \ = \ \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y}$$

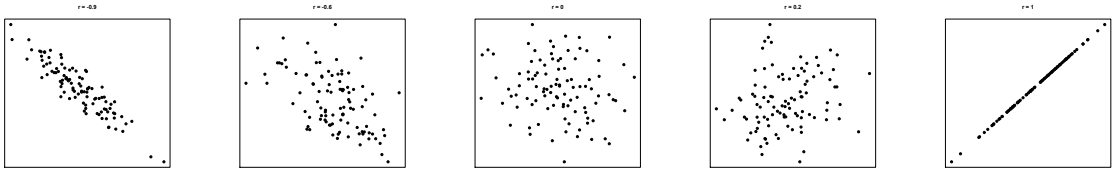
(divide by $n-1$ instead of n if this is also done for the standard deviations s_x, s_y).

Correlation measures linear association

A numerical summary of these pairs of data is given by: $\bar{x}, s_x, \bar{y}, s_y, r$.

As a convention the variable on the horizontal axis is called **explanatory variable** or **predictor**, the one on the vertical axis is called **response variable**.

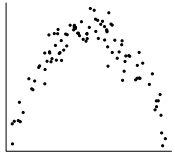
r is always between -1 and 1. The sign of r gives the direction of the association and its absolute value gives the strength:



Since both x and y were standardized when computing r, r has no units and is not affected by changing the center or the scale of either variable.

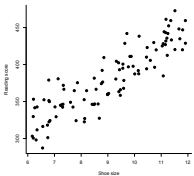
Correlation measures linear association

Keep in mind that r is only useful for measuring *linear* association:



$$r = 0$$

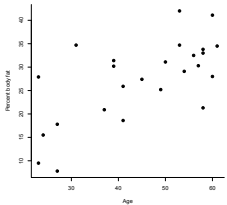
Also remember that correlation does not mean causation:

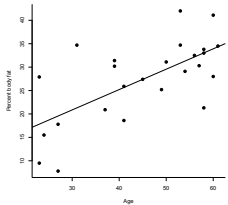


Among school children there is a high correlation between shoe size and reading ability. Both are driven by the *lurking variable* 'age'.

The regression line

If the scatterplot shows a linear association, then this relationship can be summarized by a line.





To find this line for n pairs of data $(x_1, y_1), \ldots, (x_n, y_n)$, recall that the equation of a line produces the y-value $\hat{y}_i = a + bx_i$. The idea is to choose the line that minimizes the sum of the squared distances between the observed y_i and the $\hat{y}_i$. In other words, find a and b that minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \;=\; \sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

The method of least squares

For n pairs of data $(x_1, y_1), \ldots, (x_n, y_n)$, find a and b that minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \;=\; \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

This is the **method of least squares**. It turns out that $b = r\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$. This line $\hat{y} = a + bx$ is called the **regression line**.

There is another interpretation of the regression line:
it computes the average value of y when the first coordinate is near x.

Remember that often times an average is the 'best' predictor. This shows how the regression line incorporates the information given by x to produce a good predictor of y.

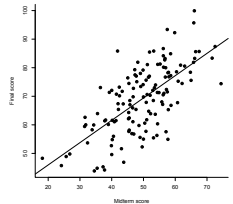
Regression to the mean

The main use of regression is to predict y from x:
Given x, predict y to be $\hat{y} = a + bx$.

The prediction for y at $x = \bar{x}$ is simply $\hat{y} = \bar{y}$.

But $b = r\frac{s_y}{s_x}$ means that if x is one standard deviation s_x above $\bar{x}$, then the predicted $\hat{y}$ is only $r\, s_y$ above $\bar{y}$.

Since r is between -1 and 1, the prediction is 'towards the mean': $\hat{y}$ is fewer standard deviations away from $\bar{y}$ than x is from $\bar{x}$.



Regression to the mean

This is called **regression to the mean** (or: the **regression effect**). In can be observed in data whose scatter is football-shaped such as the exam scores: In such a test-retest situation, the top group on the test will drop down somewhat on the retest, while the bottom group moves up.

A heuristic explanation is this: To score among the very top on the midterm requires excellent preparation as well as some luck. This luck may not be there any more on the final exam, and so we expect this group to fall back a bit.

This effect is simply a consequence of there being a scatter around the line.
Erroneously assuming that this occurs due to some action (e.g. 'the top scorers on the midterm slackened off') is the **regression fallacy**.

Predicting y from x and x from y

If we are given x, then we use the regression line $\hat{y} = a + bx$ to predict y.
To find this regression line we need only $\bar{x}, \bar{y}, s_x, s_y$ and r.

We can use software to compute this line, e.g. 'lm' in **R**, but it can also be done quickly by hand:

$\overline{\text{midterm}} = 49.5, \overline{\text{final}} = 69.1, s_{mid} = 10.2, s_{final} = 11.8, r = 0.67.$
Predict the final exam score of a student who scored 41 on the midterm.

Predict x from y

Predict the midterm score of a student who scored 89 on the final.

When predicting x from y **it is a mistake** to use the regression line $\hat{y} = a + bx$, derived for regressing y on x, and solve for x. This is because regressing x on y will result in a **different regression line**.

To avoid confusing these, always put the predictor on the x-axis and proceed as on the previous slide.

Normal approximation in regression

Regression requires that the scatter is football-shaped. Then one may use normal approximation for the y-values conditional on x. That is, the observations whose first coordinate is near that x have y-values that approximately follow the normal curve.

To standardize, subtract off the predicted value $\hat{y}$, then divide by $\sqrt{1-r^2} \times s_y$.

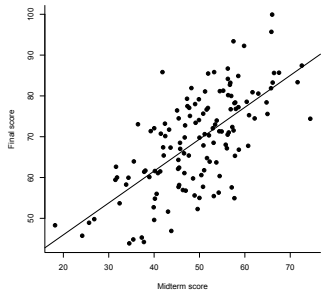
Among the students who scored around 41 on the midterm, what percentage scored above 60 on the final?

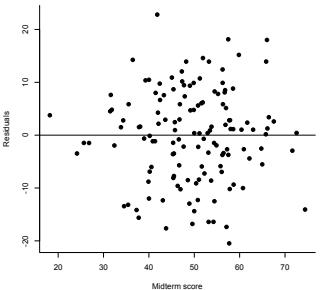
Residuals

The differences between observed and predicted y-values are called **residuals**:

$$e_i \;=\; y_i - \hat{y}_i, \qquad i = 1, \ldots, n$$

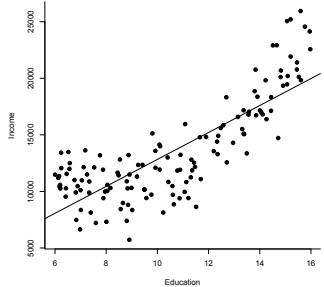
Residuals are used to check whether the use of regression is appropriate. The **residual plot** is a scatterplot of the residuals against the x-values. It should show an unstructured horizontal band.

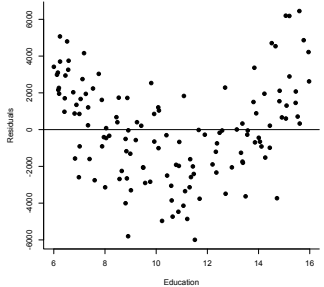




Residual plots

A curved pattern suggests that the scatter is not linear:

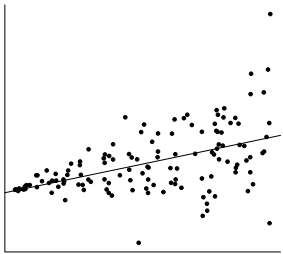


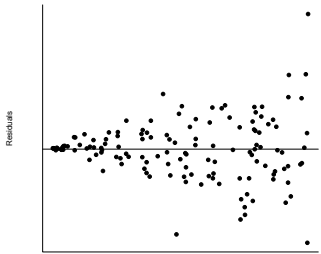


But it may still be possible to analyze these data with regression! Regression may applicable after **transforming** the data, e.g. regress $\sqrt{\text{income}}$ or log(income) on Education.

Transformations of the variables

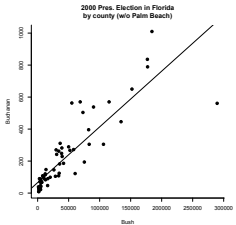
Another violation of the football-shaped assumption about the scatter arises if the scatter is **heteroscedastic**:

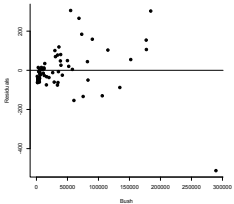




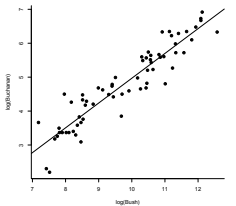
A transformation of the y-variables may produce a **homoscedastic** scatter, i.e. result in equal spread of the residuals across x. (However, it may also result in a non-linear scatter, which may require a second transformation of the x-values to fix!)

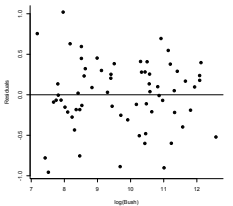
Transformation of the variables





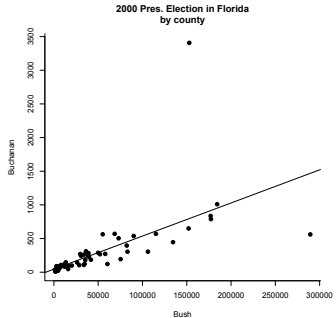
The residual plot looks heteroscedastic. Taking log of both variables produces a residual plot that is very satisfactory:

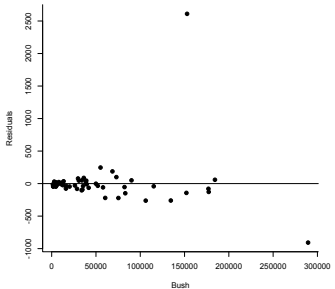




Outliers

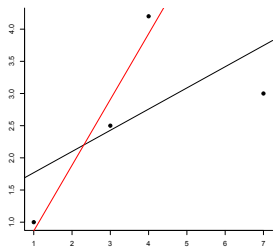
Points with very large residuals (**outliers**) should be examined: they may represent typos or interesting phenomena.





Leverage and influential points

A point whose x-value is far from the mean of the x-values has high **leverage**: it has the potential to cause a big change the regression line.



Whether it does change the line a lot ($\rightarrow$ **influential point**) or not can only be determined by refitting the regression without the point. An influential point may have a small residual (because it is influential!), so a residual plot is not helpful for this analysis.

Some other issues

- Avoid predicting y by **extrapolation**, i.e. at x-values that are outside the range of the x-values that were used for the regression: The linear relationship often breaks down outside a certain range.
- Beware of data that are summaries (e.g. averages of some data). Those are less variable than individual observations and correlations between averages tend to overstate the strength of the relationship.
- Regression analyses often report 'R-squared': $R^2 = r^2$. It gives the fraction of the variation in the y-values that is explained by the regression line. (So $1 - r^2$ is the fraction of the variation in the y-values that is left in the residuals.)

Confidence intervals

Let's look at the Gallup poll on the approval rating of the US President.
Suppose 60% of the 140 million likely voters approve of the way the president is handling his job.

Gallup polls 1,000 of them. The resulting approval percentage in the sample will be off the population percentage of 60% due to chance error. How much?
The SE tells us the likely size of the chance error.

Confidence intervals give a more precise statement.

Confidence intervals

According to the central limit theorem, the sample percentage follows the normal curve with expected value $\mu = 60\%$ and SE equal to $\frac{\sigma}{\sqrt{1000}} = \frac{0.49}{31.6} = 1.6\%$.
(Because we sample from a population of 140 million labels, of which 60% are 1s and 40% are 0s.)

By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage μ.

Confidence intervals

By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage μ.

But that is the same as saying that the population percentage is no more than 2 SEs away from the sample percentage.

So once we compute the sample percentage from our sample, say 58%, then we can give a range of plausible values for the unknown population percentage by going 2 SEs in each direction:

[54.8%, 61.2%] is a 95% **confidence interval** for the population percentage.

Interpretation of a confidence interval

[54.8%, 61.2%] is a 95% confidence interval for the population percentage.

Why 'confidence' instead of 'probability'?

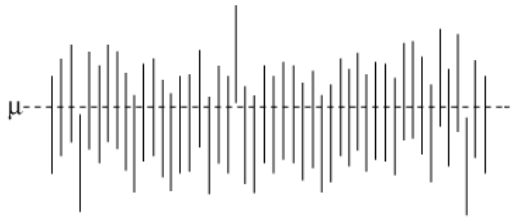
The population percentage μ is a *fixed* number, which either falls into [54.8%, 61.2%] or not. There are no chances involved.

Rather, the chances are in the sampling procedure:
a different sample of 1,000 voters will give a slightly different interval.

If one does many polls (the Gallup poll is done frequently), then 95% of these intervals trap the population percentage, and 5% will miss it.

95% is called the **confidence level**.

Interpretation of a confidence interval



"I am 95% confident that the President's approval rating is between 54.8% and 61.2%" means that 95% of the time I am correct when making such a statement based on a poll.

Keep in mind that the interval varies from sample to sample, while the population percentage is a fixed number.

Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter* μ.

Usually the confidence interval is centered at an estimate for μ which is an average.

Examples:

1. μ = approval percentage among all 140 million likely voters.
   estimate = approval percentage among voters in the sample.
2. μ = speed of light.
   estimate = average of 30 measurements.

Since the central limit theorem applies for averages, the confidence interval has a simple form:

$$\text{estimate} \pm z\,\text{SE}$$

where z is the z-score corresponding to the desired confidence level:

95% confidence level $\rightarrow z = 1.96$

90% confidence level $\rightarrow z = 1.65$

99% confidence level $\rightarrow z = 2.58$

Estimating the SE with the bootstrap principle

SE is the standard error of the estimate. If the estimate is an average (e.g. a percentage), then we know that

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population.

Now we have a problem: we don't know σ because we don't know all the data in the population (that's the reason why we sample in the first place!)

The **bootstrap principle** states that we can estimate σ by its sample version s and still get an approximately correct confidence interval.

Estimating the SE with the bootstrap principle

The **bootstrap principle** states that we can estimate σ by its sample version s and still get an approximately correct confidence interval.

Examples:

1. We poll 1,000 likely voters and find that 58% approve of the way the president handles his job.

   SE $= \frac{\sigma}{\sqrt{n}} \times 100\%$, where $\sigma = \sqrt{p(1-p)}$, p = proportion of all voters who approve.

   The bootstrap principle replaces σ by s = standard deviation of the 0/1 labels in the sample $= \sqrt{0.58(1-0.58)} = 0.49$.

   So a 95% confidence interval for p is

$$58\% \ \pm \ 2\frac{0.49}{\sqrt{1000}}, \quad \text{which is } [54.9\%, 61.1\%]$$

Estimating the SE with the bootstrap principle

2. estimate = average of 30 measurements of the speed of light.

   What is the population or probability histogram from which we sample?

   The reason we get 30 different measurementss is because each is off by a chance error:

   $$\text{measurement} = \text{speed of light} + \text{measurement error}$$

   The measurement error follows a probability histogram that is unknown to us. We estimate the standard deviation σ of this probability histogram by the standard deviation s of the sample of 30 measurements.

Later we will see how the bootstrap principle can be used to approximate the SE of estimates other than averages.

More about confidence intervals

- The width of the confidence interval is determined by z SE, which is called the **margin of error**.

  A larger sample size n will result in a smaller margin of error since SE $= \frac{\sigma}{\sqrt{n}}$, but we need *four times the sample size to cut the width in half*.

  We can also make the width smaller by making z smaller, e.g. use a 80% confidence level instead of 95%. Then the price for more precision (i.e. shorter interval) is less confidence that it covers the parameter μ.

- There is an easy to remember formula for a 95% confidence interval for a percentage:

  $$\text{estimated percentage} \pm \frac{1}{\sqrt{n}}$$

  That's because $\sigma = \sqrt{p(1-p)} \leq \frac{1}{2}$ no matter what p is.

- It is journalistic convention to use a 95% confidence level unless stated otherwise.

The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The **null hypothesis, $\mathbf{H}_0$**, states that "nothing extraordinary is going on". So in this case

$$\mathrm{H}_0\text{: } \mathrm{P(T)} = \tfrac{1}{2}$$

The **alternative hypothesis, $\mathbf{H}_A$**, states that there is a different chance process that generates the data. Here we can take

$$\mathrm{H}_A\text{: } \mathrm{P(T)} \neq \tfrac{1}{2}$$

Hypothesis testing proceeds by collecting data and evaluating whether the data are compatible with H_0 or not (in which case one **rejects $\mathbf{H}_0$**).

The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.
In this case "nothing extraordinary going on" means that the drug has no effect. So

H_0: no change in blood pressure H_A: blood pressure drops

Note that in this case the company would like to reject H_0!

So the logic of testing is typically indirect: One assumes that nothing extraordinary is happening and then hopes to reject this assumption H_0.

Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if H_0 were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

'Observed' is a statistic that is appropriate for assessing H_0. In the example of the 10 coin tosses, appropriate statistics would be the number of tails or the percent of tails.

'Expected' and SE are the expected value and the SE of this statistic, *computed under the assumption that H_0 is true.*

In the example: Using the formulas for the sum of 0/1 labels we get 'expected' $= 10 \times \frac{1}{2} = 5$ and SE $= \sqrt{10}\sqrt{\frac{1}{2} \times \frac{1}{2}} = 1.58$. So

$$z = \frac{7-5}{1.58} = 1.27$$

p-values measure the evidence agains H_0

Large values of $|z|$ are evidence against H_0: The larger $|z|$ is, the stronger the evidence. The strength of the evidence is measured by the
p-value (or: **observed significance level)**:

The p-value is the probability of getting a value of z as extreme or more extreme than the observed z, assuming H_0 is true.

But if H_0 is true, then z follows that standard normal curve, according to the central limit theorem, so the p-value can be computed with normal approximation:

The smaller the p-value, the stronger the evidence against H_0. Often the criterion for rejecting H_0 is a p-value smaller than 5%. Then the result is called **statistically significant**.

p-values measure the evidence against H_0

In the example:

Note that the p-value **does not** give the probability that H_0 is true, as H_0 is either true or not - there are no chances involved. Rather, it gives the probability of seeing a statistic as extreme, or more extreme, that the observed one, assuming H_0 is true.

Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the conents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

To write this down formally we introduce 0/1 labels since we are counting correct answers: 1 = correct answer, 0 = wrong answer

H_0: $P(0) = P(1) = \frac{1}{2}$ H_A: $P(1) > \frac{1}{2}$

This is a **one-sided test**: the alternative hypothesis for for $P(1)$ we are interested in is on one side of $\frac{1}{2}$.

Distinguishing Coke and Pepsi by taste

Since we are looking at the sum of ten 0/1 labels, the z-statistic is the same that we had for coin-tossing:

$$z = \frac{\text{observed sum} - \text{expected sum}}{\text{SE of sum}} = \frac{7-5}{1.58} = 1.27$$

But since we do a one-sided test instead of a two-sided test, the p-value is only half as large:

Since 10.2% is not smaller than 5%, we don't reject H_0: We are not convinced that the student can distinguish Coke and Pepsi.

Distinguishing Coke and Pepsi

A two-sided alternative might also be appropriate:

H_A: $P(1) \neq \frac{1}{2}$

H_A corresponds to a student who is more likely than not to distinguish Coke and Pepsi, but who may confuse them. Such a student might get one correct answer (say).

One has to carefully consider whether the alternative should be one-sided or two-sided, as the p-value gets doubled in the latter case.

It is not ok to change the alternative afterwards in order to get the p-value below 5%.

The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration μ in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

So it may be that the concentration μ is below 15 ppb, but measurement error results in an average of 15.6 ppb.

H_0: $\mu = 15$ ppb $\qquad$ H_A: $\mu > 15$ ppb

We can try a z-test for the average of the measurements:

$$z = \frac{\text{observed average} - \text{expected average}}{\text{SE of average}} = \frac{15.6 \text{ ppb} - 15 \text{ ppb}}{\text{SE of average}}$$

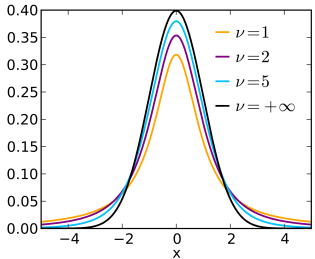
since the measurement error has expected value zero.

The t-test

SE of average $= \frac{\sigma}{\sqrt{n}}$, but the standard deviation σ of the measurement error is unknown.

We can estimate σ by s, the sample standard deviation of the measurements. However:

If we estimate σ and n is small ($n \leq 20$), then the normal curve is not a good enough approximation to the distribution of the z-statistic. Rather, an appropriate approximation is **Student's t-distribution with $n-1$ degrees of freedom**:



The t-test

The fatter tails account for the additional uncertainty introduced by estimating σ by $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Using the **t-test** in place of the z-test is only necessary for small samples: $n \leq 20$ (say).

In that case it is also better to replace the confidence interval $\bar{x} \pm z$ SE by

$$\bar{x} \pm t_{n-1}\text{SE}$$

More on testing

- *Statistically significant does not mean that the effect size is important*:

  Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.

  That may not be of practical concern, even though the test may be highly signficant: Statistical significance convinces us that there is an effect, but it doesn't say how big the effect is.

  Reason: A large sample size n makes SE $= \frac{\sigma}{\sqrt{n}}$ small, so even a small exceedance over the limit by (say) 0.05 ppb may give a statistically significant result.

  Therefore it is helpful to complement a test with a confidence interval: In the above case a 95% confidence interval for μ might be [15.02 ppb, 15.08 ppb].

More on testing

- There is a general connection between confidence intervals and tests:

  A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.
  (A 5% **significance level** means that the threshold for the p-value is 5%).
- There are two ways that a test can result in a wrong decision:

  H_0 is true, but was erroneously rejected $\rightarrow$ Type I error ('false positive')

  H_0 is false, but we fail to reject it $\rightarrow$ Type II error

  Rejecting H_0 if the p-value is smaller than 5% means P(type I error) $\leq 5\%$

The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

p_1 =proportion of all likely voters approving last month

is equal to

p_2 =proportion of all likely voters approving this month

"nothing unusual is going on" means $p_1 = p_2$. It's common to look at the difference $p_2 - p_1$ instead:

$$\mathsf{H}_0 : \ p_2 - p_1 = 0 \qquad\qquad \mathsf{H}_1 : \ p_2 - p_1 \neq 0$$

p_1 is estimated by $\hat{p}_1 = 55\%$, p_2 by $\hat{p}_2 = 58\%$. The central limit theorem applies to the difference $\hat{p}_2 - \hat{p}_1$ just as it does to $\hat{p}_1$ and $\hat{p}_2$. So we can use a z-test:

The two-sample z-test

We can use a z-test for the difference $\hat{p}_2 - \hat{p}_1$:

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if $\hat{p}_1$ and $\hat{p}_2$ are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}. \qquad \text{So}$$

$$z = \frac{(\hat{p}_2 - \hat{p}_1) - 0}{\sqrt{\sqrt{\frac{p_1(1-p_1)}{1000}}^2 + \sqrt{\frac{p_2(1-p_2)}{1500}}^2}} = \frac{0.03}{0.0202} = 1.48$$

The two-sample z-test

The confidence interval for $p_2 - p_1$ is

$$(\hat{p}_2 - \hat{p}_1) \ \pm \ z \ \mathsf{SE}(\hat{p}_2 - \hat{p}_1) \ = \ [-1\%, 7\%] \quad \text{when } z = 2$$

We can improve the estimate of $\mathsf{SE}(\hat{p}_2 - \hat{p}_1)$ somewhat by using the fact that $p_1 = p_2$ on H_0. Since there is a common proportion we can estimate it by **pooling** the samples:

$0.55 \times 1000 = 550$ voters approve in the first sample, 870 in the second, so in total there are 1420 approvals out of 2500. So the **pooled estimate** of $p_1 = p_2$ is $\frac{1420}{2500} = 56.8\%$.

So we estimate $\mathsf{SE}(\hat{p}_2 - \hat{p}_1)$ by $\sqrt{\frac{0.568(1-0.568)}{1000} + \frac{0.568(1-0.568)}{1500}} = 0.02022$, which essentially gives the same answer in this case.

The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$\mathsf{SE}(\bar{x}_2 - \bar{x}_1) \;=\; \sqrt{(\mathsf{SE}(\bar{x}_1))^2 + (\mathsf{SE}(\bar{x}_2))^2}$$

and $\mathsf{SE}(\bar{x}_1) = \frac{\sigma_1}{\sqrt{n_1}}$ is estimated by $\frac{s_1}{\sqrt{n_1}}$.

If the sample sizes n_1, n_2 are not large, then the p-value needs to computed from the t-distribution.

The pooled standard deviation

If one has reason to assume that $\sigma_1 = \sigma_2$ (or if this has been checked), then one may use the **pooled estimate** for $\sigma_1 = \sigma_2$ given by

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

However, the advantages of using s^2_{pooled} are small and the analysis rests on the assumption that $\sigma_1 = \sigma_2$. For these reasons the pooled t-test is usually avoided.

All of the above two-sample tests require that the two samples are independent. They are also applicable in special situations where the samples are dependent, e.g. to compare the treatment effect when subjects are randomized into treatment and control groups.

The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

The two-sample t-test is not applicable since the two samples are not independent. Even if they were independent, the small differences in ages would not be significant since the standard deviations are large for husbands and also for the wives.

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

H_0: population difference is zero

$t = \frac{\bar{d}-0}{\mathrm{SE}(\bar{d})}$, where d_i is the age difference of the ith couple.

$\mathrm{SE}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}$. Estimate σ_d by $s_d = 0.55$. Then $t = \frac{1.4-0}{0.55/\sqrt{5}} = 5.69$

The independence assumption is in the sampling of the couples.

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

We can test

H_0: half the husbands in the population are older than their wives

using 0/1 labels and a z-test, just as we tested whether a coin is fair:

$$z = \frac{\text{sum of 1s} - \frac{n}{2}}{\text{SE of sum}} = \frac{5 - \frac{5}{2}}{\sqrt{5}\frac{1}{2}} = 2.24 \quad \text{since } \sigma = \tfrac{1}{2} \text{ on } H_0.$$

The p-value of this **sign-test** is less significant than that of the paired t-test. This is because the latter uses more information, namely the size of the differences. On the other hand, the sign test has the virtue of easy interpretation due to the analogy to coin tossing.

Simulations in statistical inference

Recall the confidence interval for the population mean:

$$\bar{x} \pm z\,\mathsf{SE}(\bar{x})$$

The idea was that $\bar{x}$ follows approximately the normal curve.

What if we are interested in an estimator $\hat{\theta}$ for some parameter θ and the normal approximation is not valid for the estimator $\hat{\theta}$? What if there is no formula for $\mathsf{SE}(\hat{\theta})$?

In such situations, simulations can often be used to estimate these quantities quite well. In fact, simulations may result in better estimates even in cases where the normal approximation is applicable!

The Monte Carlo Method

What is the average height of all people living in the United States?

This is difficult to determine exactly but can easily be estimated quite well:

Sample $n = 100$ (say) people at random. Then use the average height of these n people as an estimate of the average height of all people in the US.

This is an example of the general problem where we are interested in a unknown **parameter** θ of a population.

We estimate θ with a **statistic** (**estimator**) $\hat{\theta}$ which is based on a sample of n observations $X_1, \ldots, X_n$ drawn at random from the population:

$\hat{\theta}$ = average of the sample $= \frac{1}{n} \sum_{i=1}^{n} X_i$

The Monte Carlo Method

$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ tends to be close to the uncomputable population mean θ, even for moderate sample sizes such as $n = 100$.

This example is a special case of the **Monte Carlo Method** or **Simulation**:

- We approximate a fixed quantity θ by the average of independent random variables that have expected value θ.
- By the law of large numbers, the approximation error can be made arbitrarily small by using a large enough sample size.

The Monte Carlo Method

The Monte Carlo Method can also be used for more involved quantities. For example, we can use it to compute the standard error (SE) of a statistic $\hat{\theta}$.

Recall that the standard error tells roughly how far off the statistic will be from its expected value. The precise definition is

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{E}(\hat{\theta} - E(\hat{\theta}))^2}.$$

- Get many (say 1,000) samples of 100 observations each.
- Compute $\hat{\theta}$ for each sample, resulting in 1,000 estimates $\hat{\theta}_1, \ldots, \hat{\theta}_{1000}$.
- Compute the standard deviation of these 1,000 estimates:
  $s(\hat{\theta}_1, \ldots, \hat{\theta}_{1000}) = \sqrt{\frac{1}{999} \sum_{i=1}^{1000} (\hat{\theta}_i - \mathrm{ave}(\hat{\theta}_i))^2}$

Note that this is not an average of independent random variables. But it can be shown that the law of large numbers still applies and Monte Carlo works:
$s(\hat{\theta}_1, \ldots, \hat{\theta}_{1000}) \approx \mathrm{SE}(\hat{\theta})$.

We can use Monte Carlo only if we can draw many samples of size 100!

The Bootstrap principle

We have an estimate $\hat{\theta}$ for a parameter θ and want to know how accurate $\hat{\theta}$ is:
we would like to find $\mathrm{SE}(\hat{\theta})$ or give a confidence interval for θ.

The bootstrap can do this in quite general settings.

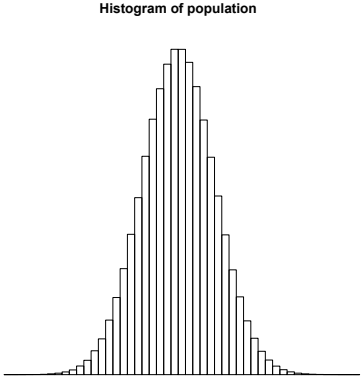
Example: θ = average height of all people in the US.
θ is unknown but can be estimated by the average height $\hat{\theta}$ of 100 randomly selected people.

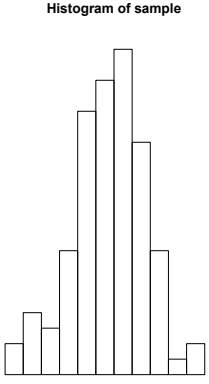
This illustrates the **plug-in principle**:

We can't compute the population mean because we can't access the whole population. So we 'plug in' the sample in place of the population and compute the mean of the sample instead.

The Bootstrap principle

The rationale for the plug-in principle is that the sample mean $\hat{\theta}$ will be close to the population mean θ because the sample histogram is close to the population histogram.





The bootstrap principle

The bootstrap uses the plug-in principle and the Monte Carlo Method to approximate quantities such as SE($\hat{\theta}$).

Here is the reasoning behind the bootstrap:

Suppose we can draw as many samples from the population as we wish. Then we can approximate SE($\hat{\theta}$) with Monte Carlo:

- Draw a sample $X_1, \ldots, X_n$ and use it to compute $\hat{\theta}$.
- Repeat B times (say B=1,000) to get $\hat{\theta}_1, \ldots, \hat{\theta}_B$.
- The standard deviation of these B estimates is close to SE($\hat{\theta}$) if B is large, by the law of large numbers.

However, we have only one sample $X_1, \ldots, X_n$ and we can't simulate more because the population is not accessible.

The bootstrap uses the plug-in principle to get around this: It simulates from the sample instead of from the population.

The bootstrap principle

The bootstrap pretends that the sample histogram is the population histogram and then uses Monte Carlo to simulate the quantity of interest.

Simulating a bootstrap sample $X_1^*, \ldots, X_n^*$ means that we draw n times with replacement from $X_1, \ldots, X_n$.

The bootstrap consists of two steps:

- Draw B bootstrap samples and compute $\hat{\theta}^*$ for each bootstrap sample:
  $$X_1^{*1}, \ldots, X_n^{*1} \;\rightarrow\; \hat{\theta}_1^*$$
  $$\vdots$$
  $$X_1^{*B}, \ldots, X_n^{*B} \;\rightarrow\; \hat{\theta}_B^*$$
- Use $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ to approximate the quantity of interest.
  For example, we approximate $\mathrm{SE}(\hat{\theta})$ by the standard deviation of $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

More about the bootstrap

The **nonparametric bootstrap** simulates a bootstrap sample $X_1^*, \ldots, X_n^*$ by drawing with replacement from $X_1, \ldots, X_n$.

Sometimes a parametric model is appropriate for the data, e.g. a normal distribution with unknown mean and standard deviation. Then one may be better off with the **parametric bootstrap**, which simulates the bootstrap samples from this model, using estimates for the unknown parameters.

So far, the bootstrap samples were drawn independently. If there is dependence in the data (time series), then this needs to be incorporated, e.g. with the **block bootstrap**.

Bootstrap confidence intervals

If the sampling distribution of $\hat{\theta}$ is approximately normal, then

$$\left[\hat{\theta} - z_{\alpha/2}\,\mathrm{SE}(\hat{\theta}), \hat{\theta} + z_{\alpha/2}\,\mathrm{SE}(\hat{\theta})\right]$$

is an approximate $(1-\alpha)$-confidence interval for θ.

$\mathrm{SE}(\hat{\theta})$ can be estimated by the bootstrap.

If $\hat{\theta}$ is far from normal, then we have to use the bootstrap to estimate the whole sampling distribution of $\hat{\theta}$, not just $\mathrm{SE}(\hat{\theta})$.

The sampling distribution of $\hat{\theta}$ can be approximated by that of $\hat{\theta}^*$, which in turn can be approximated by the histogram of $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.

This gives the **bootstrap percentile interval**

$$\left[\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}\right]$$

where $\hat{\theta}^*_{(\alpha/2)}$ is the $\alpha/2$ percentile of the $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.

Bootstrap confidence intervals

An alternative to bootstrapping the distribution of $\hat{\theta}$ is to do so for $\hat{\theta} - \theta$.

The hope is that this approach is less sensitive to θ and therefore produces a more accurate confidence interval.

This results in the **bootstrap pivotal interval**

$$\left[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\right]$$

Bootstrapping for regression

We have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the simple linear regression model

$$Y_i = a + bX_i + e_i$$

From the data we can compute estimates $\hat{a}, \hat{b}$. How can we use the bootstrap to get standard errors and confidence intervals?

- Compute the residuals $\hat{e}_i = Y_i - \hat{a} - \hat{b}X_i$
- Resample from those residuals to get $e_1^*, \ldots, e_n^*$
- Compute the bootstrapped responses $Y_i^* = \hat{a} + \hat{b}X_i + e_i^*$

This gives a bootstrap sample $(X_1, Y_1^*), \ldots, (X_n, Y_n^*)$, from which we can estimate the parameters $\hat{a}^*$ and $\hat{b}^*$ in the usual way.

Categorical data

In 1912 the Titanic sank and more than 1,500 of the 2,229 people on board died. Did the chances of survival depend on the ticket class?

	First	Second	Third	Crew
Survived	202	118	178	215
Died	123	167	528	698

This is an example of **categorical data**: the data are counts for a fixed number of categories. Here the data are tabulated in a 2×4 **table.**

Such a table is called a **contingency table** because it shows the survival counts for each category of ticket class, i.e. contingent on ticket class.

We will look at three problems involving categorical data: testing goodness-of-fit, testing homogeneity, and testing independence.

Testing goodness-of-fit

In 2008 the producer of M&Ms stopped publishing the color distribution of M&Ms. Here are the last percentages published in 2008:

Blue	Orange	Green	Yellow	Red	Brown
24%	20%	16%	14%	13%	13%

Recently, students in my class inspected the contents of one bag of milk chocolate M&Ms and got the following counts:

Blue	Orange	Green	Yellow	Red	Brown
85	79	56	64	58	68

Are these counts consistent with the last published percentages? Or is there sufficient evidence to claim that the color distribution is now different?

This question requires a **test of goodness-of-fit** for the six categories.

Testing goodness-of-fit

H_0 = "nothing is extraordinary is going on" = the color distribution is given by the table from 2008.

H_A: the color distribution is different from the 2008 table

The idea is to compare the observed counts to the numbers one would expect if H_0 is true.

To compute the expected counts, note that we have a total of 410 M&Ms. So on H_0 we expect $24\% \times 410 = 98.4$ blue M&Ms. After doing this computation for the other colors, we get a table of observed and expected counts:

	Blue	Orange	Green	Yellow	Red	Brown
Observed	85	79	56	64	58	68
Expected	98.4	82	65.6	57.4	53.3	53.3

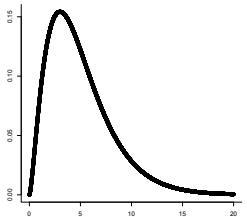
Testing goodness-of-fit

Next we combine the differences between observed and expected counts across all six categories:

$$\chi^2 = \sum_{\text{all categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$= \frac{(85 - 98.4)^2}{98.4} + \frac{(79 - 82)^2}{82} + \ldots + \frac{(68 - 53.3)^2}{53.3} = 8.57$$

Large values of the **chi-square statistic** χ^2 are evidence against H_0. The p-value is the *right* tail of the χ^2 distribution with degrees of freedom = number of categories$-1 = 5$.



Testing goodness-of-fit

This χ^2-test is applicable if the observed data are counts that were obtained by drawing the M&Ms independently from a population of M&Ms - it is plausible that filling the M&Ms into bags follows approximately this process.

Blue M&Ms were not introduced until 1995. Initially some people suspected that there were fewer blue M&Ms in a bag than advertised. Note that testing the proportion of one category can be done with the z-test. The chi-square test provides an extension of the z-test to testing several categories.

Testing homogeneity

Are the color distributions the same for milk chocolate M&Ms, peanut M&Ms, and caramel M&Ms?

The χ^2-**test of homogeneity** tests the null hypothesis that the distribution of a categorical variable (color) is the same for several populations (milk, peanut, caramel).

It assumes that the samples are drawn independently within and across the populations.

To see how the test works, let's look at the survival data for the Titanic:

	First	Second	Third	Crew
Survived	202	118	178	215
Died	123	167	528	698

Testing homogeneity

Note that in this case we are not sampling from a population: The data are not a random sample of the people on board, rather the data represent the whole population.

So in this case the chance process resulting in survival or death is not in the sampling, but the result of the random events occuring when looking for a way out of the ship, getting into a life boat or into the water, being rescued out of the water in time etc.

Then the 325 observations about first class passengers represent 325 independent draws from a probability histogram that gives a certain chance for survival. The 285 observations about second class passengers are drawn from the probability histogram for second class passengers, which may be different. The null hypothesis says that the probability of survival is the same for all four probability histograms.

Testing homogeneity

	First	Second	Third	Crew
Survived	202	118	178	215
Died	123	167	528	698

In order to compute the table of expected entries, note that under H_0 the probability of survival is the same for all four populations. So we can estimate this probability by pooling the data: in total there were 713 survivors among the 2,229 people on board, so we estimate the probability of survival as $\frac{713}{2229} = 32\%$.

So the expected number of surviving first class passengers is $32\% \times 325 = 104.0$. Doing this computation for the other seven categories gives the table of expected values:

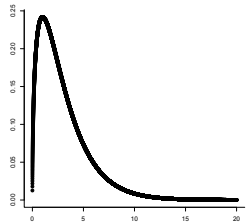
	First	Second	Third	Crew
Survived	104.0	91.2	225.8	292.1
Died	221.0	193.8	480.1	620.8

Testing homogeneity

Now we can compute the chi-square statistic over all 8 cells in the table:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$
$$= \frac{(202 - 104)^2}{104} + \ldots = 192.2$$

In this case the degrees of freedom
$= (\text{no. of columns} - 1) \times (\text{no. of rows} - 1) =$
$(4 - 1) \times (2 - 1) = 3.$



Testing independence

Is gender (male/female) related to voting preference (liberal/conservative)?

Now we have two categorical variables: gender and voting preference. The null hypothesis is that the two variables are independent. The alternative hypothesis is that there is some kind of association.

The chi-square test can be used to test this null hypothesis:
After sampling from the population the counts are recorded in a 2×2 table. The χ^2-statistic and p-value are computed exactly as in the case of testing homogeneity.

Comparing these two uses of the χ^2-test:

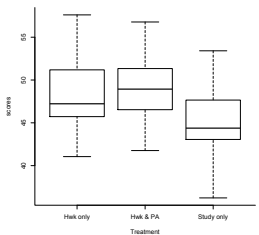
	Sample	Research question
χ^2-test of homogeneity	single categorical variable measured on several samples	Are the groups homogeneous (have the same distribution of the categorical variable)?
χ^2-test of independence	two categorical variables measured on one sample	Are the two categorical variables independent?

Comparing several groups of data

Does peer assessment (i.e. a student is grading other students' homework) enhance learning?

Students were randomized into three groups, each spending the same amount of time per week on either: homework only, homework and peer assessment, studying without doing homework.

The final exam scores were summarized by boxplots:



Is there sufficent evidence to conclude that the three treatments result in different outcomes?

Comparing several groups of data

H_0 = "nothing extraordinary is going on" = all group means are equal.

Recall that if we have only two groups, then we can compare them with the two-sample t-test:

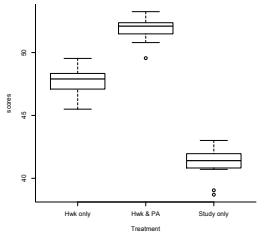
$$t = \frac{\text{difference in sample means}}{\text{SE of difference}}$$

t compares the size of the difference in sample means to the size of the chance variability as measured by the SE.

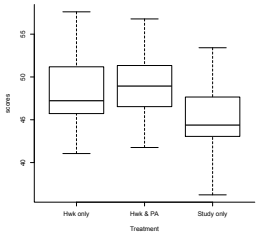
ANOVA generalizes this idea.

Analysis of Variance

Consider these two hypothetical outcomes:



The differences between the sample means are large relative to the variability within the groups, suggesting that there is a difference in treatment means.



The differences between the sample means are small compared to the variability within the groups: It might well be due to chance variability.

Analysis of Variance

The key idea to make this precise is to compare the sample variance of the means to the sample variance within the groups.

That's why this methodology is called Analysis of Variance (ANOVA).

But recall that according to the square root law, the chance variability in the sample mean is smaller than the chance variability in the data. So the evidence against H_0 is not obvious from the boxplots. Rather, a computation is necessary.

Analysis of Variance

We have k groups and the jth group has n_j observations:

group 1	group 2	...	group k
y_{11}	y_{12}		y_{1k}
$\vdots$	$\vdots$		$\vdots$
$y_{n_1 1}$	$y_{n_2 2}$		$y_{n_k k}$

In total there are $N = n_1 + \ldots + n_k$ observations.

The sample mean of the jth group is $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$.

The overall sample mean is $\bar{\bar{y}} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}$.

Analysis of Variance

The **treatment sum of squares**

$$\mathsf{SST} = \sum_j \sum_i (\bar{y}_j - \bar{\bar{y}})^2$$

has $k-1$ degrees of freedom.The **treatment mean square**

$$\mathsf{MST} = \frac{\mathsf{SST}}{k-1}$$

measures the variability of the treatment means $\bar{y}_j$. The **error sum of squares**

$$\mathsf{SSE} = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$$

has $N-k$ degrees of freedom.The **error mean square**

$$\mathsf{MSE} = \frac{\mathsf{SSE}}{N-k}$$

measures the variability within the groups.

The Analysis of Variance F-test

Since we want to compare the variation between the groups to the variation within the groups we look at the ratio

$$F = \frac{\text{MST}}{\text{MSE}}$$

Under the null hypothesis of equal group means this ratio should be about 1. It will not be exactly 1 due to sampling variability:

It follows a F-**distribution with** $k-1$ **and** $N-k$ **degrees of freedom.**

Large values of F suggest that the variation between the groups is unusually large. We reject H_0 if F is in the right 5% tail, i.e. when the p-value is smaller than 5%.

The ANOVA table

All the relevant information is summarized in the **ANOVA table**:

Source	df	Sum of Squares	Mean Square	F	p-value
Treatment	$k-1$	SST	MST	MST/MSE	
Error	$N-k$	SSE	MSE		
Total	$N-1$	TSS			

where TSS $= \sum_j \sum_i (y_{ij} - \bar{\bar{y}})^2$.

The ANOVA table

For the data of the boxplots in the right display we get a p-value of 0.097, so there is not enough evidence to reject H_0:

Source	df	Sum of Squares	Mean Square	F	p-value
Treatment	2	98.4	47.2	2.49	0.097
Error	38	723.8	19.1		
Total	40	822.2			

The one-way ANOVA model

The **one-way ANOVA** model behind this table is

$$y_{ij} \;=\; \mu_j + \epsilon_{ij}$$

where μ_j is the mean of the jth group and the ϵ_{ij} are independent random variables (e.g. measurement error) that follow the normal curve with mean 0 and common variance σ^2.

So the null hypothesis is $\mu_1 = \mu_2 = \ldots = \mu_k$.

Instead of looking at the group means μ_j it is helpful to look at the deviations τ_j from the overall mean μ: $\tau_j = \mu_j - \mu$.

So the model is

$$y_{ij} \;=\; \mu + \tau_j + \epsilon_{ij}$$

where τ_j is called the **treatment effect** of group j. Then the null hypothesis becomes

$\mathrm{H}_0 :\ \tau_1 = \tau_2 = \ldots = \tau_k = 0.$

The one-way ANOVA model

We estimate the overall mean μ by the 'grand mean' $\bar{\bar{y}}$. Then the estimate of $\tau_j = \mu_j - \mu$ becomes $\bar{y}_j - \bar{\bar{y}}$. The estimate of ϵ_{ij} is the residual $y_{ij} - \bar{y}_j$.

Corresponding to the model $y_{ij} = \mu + \tau_j + \epsilon_{ij}$ we can write y_{ij} as the sum of the corresponding estimates:

$$y_{ij} = \bar{\bar{y}} + (\bar{y}_j - \bar{\bar{y}}) + (y_{ij} - \bar{y}_j)$$

It turns out that such a decomposition is also true for the sum of squares:

$$\sum_j \sum_i (y_{ij} - \bar{\bar{y}})^2 = \sum_j \sum_i (\bar{y}_j - \bar{\bar{y}})^2 + \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$$

$$\text{TSS} = \text{SST} + \text{SSE}$$

This splits the total variation TSS into two 'sources': SST and SSE.

More on ANOVA

- The F-test assumes that all the groups have the same variance σ^2. This can be roughly checked with side-by-side boxplots, and there are also formal tests.
- Another assumption was that the data are indendent within and across groups. This would be the case if the subjects were assigned to treatments at random.
- If the F-test rejects, then we can conclude that the group means are not all equal, but how do they differ?

  We can examine all pairs of means with a two-sample t-test using $s_{pooled} = \sqrt{\text{MSE}}$.

  But since that involves several tests, an adjustment such as the Bonferroni adjustment is necessary, see the next module.

Data snooping and the multiple testing fallacy

In 1992 a Swedish study examined whether living near a power line causes adverse health effects. It reported a statistically highly significant increase in childhood leukemia.

A careful follow-up analysis failed to confirm this result.

Why did the study find a statistically significant result?

The study looked at 800 different health effects:

There were 800 statistical tests involved.

Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:
A smaller p-value means stronger evidence.

p-value $< 1\% \ \rightarrow$ test is 'highly significant'

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

So if we do 800 tests, then even if there is no effect at all we expect to see $800 \times 1\% = 8$ highly significant results just by chance!

This is called the **multiple testing fallacy** or **look-elsewhere effect**.

When analyzing large amounts of data it is easy to fall into this trap because there are so many potential relationships to explore, which leads to **data snooping** (=**data dredging**).

Reproducibility and Replicability

Data snooping and other problems have lead to a crisis with regard to **replicability** (getting similar conclusions with different samples, procedures and data analysis methods) and **reproducibility** (getting the same results when using the same data and methods of analysis.)

- 'How science goes wrong' in The Economist (10/13/2013)
- 'Why most published research findings are false' by J. Ioannidis (2005)

How can one account for multiple testing?

Bonferroni correction: If there are m tests, multiply the p-values by m.

The Bonferroni correction makes sure that P(any of the m tests rejects in error) $\leq 5\%$.

The Bonferroni correction is often very restrictive: It guards against having even one false positive among the m tests.

As a consequence the adjusted p-values may not be significant any more even if a noticeable effect is present.

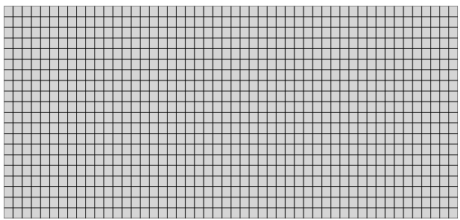
Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



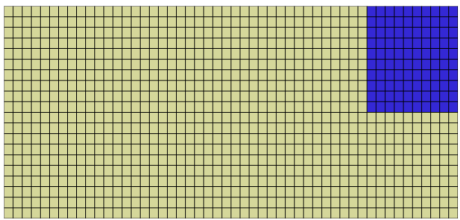
Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



In 900 cases the null hypothesis is true ("Nothing is going on"), and in 100 cases an alternative hypothesis is true ("There is an effect: something is going on").

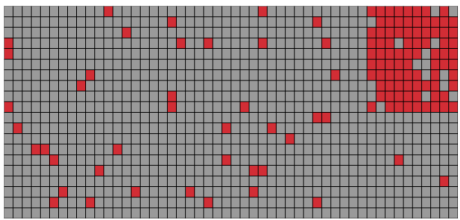
Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



Doing 1,000 tests results in Discoveries and Non-discoveries.

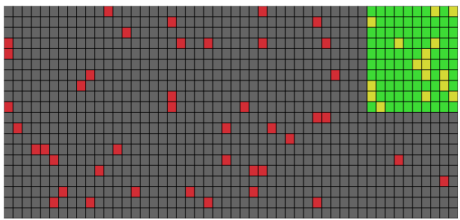
Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



We made 80 true discoveries and 41 false discoveries. The false discovery proportion is 41/121=0.34.

Accounting for multiple testing with FDR

False discovery rate (FDR): Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level $\alpha = 5\%$ (say):

1. Sort the p-values: $p_{(1)} \leq \ldots \leq p_{(m)}$
2. Find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$
3. Declare discoveries for all tests i from 1 to k.

Accounting for multiple testing with validation set

Using a validation set: Split the data into a *model-building set* and a *validation set* before the analysis.

You may use data snooping on the model-building set to find something interesting.

Then test this hypothesis on the validation set.

This approach requires strict discipline: You are not allowed to look at the validation set during the exploratory step!